MS109

Micromedia Limited

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

This statement was filmed with this document.
If the pages of the document are less clear
than this statement it is due to the poor
photographic quality of the document.

NATIONAL MICROGRAPHICS ASSOCIATION

24x

Statistics Canada | Statistique Canada

Research Development and Analysis Group

Section de la recherche du développement et de l'analyse

R-018E
C.2

TRANSFORMING MONTHLY UNIFORM CRIME REPORTS TO INCIDENT RECORDS

Research Paper No. 1

by

Craig McKie

The responsibility for the analysis and interpretation of the data is that of the author and not of Statistics Canada

Institutions and Agriculture Statistics Branch
Research, Development and Analysis Group
Statistics Canada

1981

Version française de cette publication disponible sur demande

There is a widespread problem in administrative record keeping which only becomes manifest in the subsequent interpretation of summary data files. This problem stems from the fact that organizations, in accumulating records on various phenomena, may inadvertently create for themselves types of data structures which are difficult to manipulate. Programmers have become familiar with rectangular data matrixes and the ways of manipulating them; but the problems associated with manipulating non-rectangular matrixes (which are common in administrative records where summary totals are kept) have restricted their use to a considerable extent. In such cases, what often happens is that data are perceived to be intractable for any purpose beyond elementary description. Somewhat more infrequently, expensive made-to-order systems are developed to transform such data. Such systems are costly, may require knowledge of a programming language, may be highly installation-dependant, and may not be documented to such a degree that they will survive their creator's departure or be readily modified.

This paper describes a technique based on the Statistical Package for the Social Sciences which overcomes these problems in a generalized approach to the manipulation of non-rectangular data matrixes such as are contained on administrative summary records. We first developed this technique in the Justice Statistics Division of Statistics

Canada, specifically to extend the range of possible statistical manipulation of aggregate data on crime and law enforcement produced by the Uniform Crime Reporting Program. Thus, we use as our example data from the Uniform Crime Reporting Programme.

Datasets such as that which results from the Uniform Crime Reporting Program, in which monthly totals on a large number of types of criminal incidents are submitted for electronic data capture, are inherently limited, and problem-prone for anything but rudimentary tabular analysis. These problems stem from the periodic and summary nature of the data collected.

Uniform Crime Reporting Program Data

UCR data are collected in a number of jurisdictions (notably by Statistics Canada; and by the FBI in the United States) from a large number of reporting units, generally corresponding to municipal, regional and provincial police force areas (about 1800 in Canada). These records are "the only source of data on the magnitude and trends of crime .... and as such, they are heavily relied upon by administrators, politicians, policy and opinion makers, the press, and the public at large" (Inciardi, 1978:3). However, owing to

their structure, and the lack of quality control in their preparation,* information reports based on these data are widely believed to be "incomplete and structurally biased, resulting in the creation and persistence of many myths about crime" (1978:3). While we can do nothing here about the latter type of problem** (haphazard, biased, or otherwise erroneous reporting), the former problem of statistical processing limitations is amenable to a technical fix.

To date, analysis of these data have been severely constrained by the aggregated nature of the data. Totals are collected and collated - no attempt is made to keep records on individual incidents or offenders (if their identity is even known). Rather, summed incidents of various types which are reported to, and subsequently by, police are totalled, often on a monthly basis and forwarded for

* It is widely believed that many crimes are never reported to police -- the so-called "dark figure", perhaps as high as 5 cases for every one reported.

* Substantively, the data from the UCR program are fraught with problems. They are weak with respect to: completeness of reporting; the accuracy, reliability, and consistency of classification of offences; and the application of the "most serious offence" rule which results in the non-appearance in the counts of less serious but related charges laid in conjunction with the reported offences.

aggregation to national totals. Simple distributive tabulations are the sole result -- and we can find few instances in the Canadian or American literature where more imaginative analytical steps have been taken with this type of data.

The problem therefore is to recast the UCR data in such a way that more rigorous analysis becomes possible. In so doing, we are only addressing the question of how better to process and manipulate the data as they now exist.

The Problem

The problem presented by Canadian UCR administrative data is that it is kept as a giant agglomeration of monthly tables reporting summary numbers in a large number of offence categories (132). Thus, each record represents the activity totals in a reporting unit in each of 132 types of offences for one month.

The following text describes how this data is restructured and transformed to produce an SPSS "SAVE FILE" which can produce standard annual tables.

UCR - SPSS Hybridization

Perhaps the best way of expressing in words the operation of this programme (and one rapidly exhausts the conventional vocabulary in this endeavour) is to say that it works as if one were operating on source microdata with SPSS. Two other ways of understanding it would be: a) the data are tractable with SPSS in a fairly conventional manner once the intermediate processing is completed, or b) the data structure is rearranged so that for all intents and purposes it behaves as would microdata (were there any).

This being the case, it is, strictly speaking, misleading to refer to the creation of microdata where theretofore only aggregate data existed. Rather, we suspect, one should ask why one should use SPSS in this manner to process these numbers. There are a number of such reasons:

1. SPSS is approaching the status of the most universally applied packaged programme. The product would thus be much more amenable to manipulation by novices, something which cannot be said of the current "system" which employs a number of separate made-to-measure programmes which are extremely limited in scope.

2. Using SPSS does not create strictu sensu new microdata. Rather, it rearranges the data with a view to fully exploiting the level of microdata that is implicit in this and other aggregate summary records which are stored in the form of a matrix. (This is discussed more fully below).

3. As is the case in all subject matter fields in social statistics, the study of criminology is rapidly becoming more sophisticated, with the result that a wider variety of multivariate statistical procedures are being applied to crime data. Some of these applications have been based on UCR data such as: factor structures among offence categories (Oosthoek, 1978), and trend analysis (Inciardi, 1978,) - the latter having a logical extension to the analysis of time series.* In light of these recent approaches, the need to have the capability of manipulating the UCR data comes to the fore, especially when one must perforce employ statistical packages (from the point of view of practicality) to use these procedures.

* For other Canadian examples, see also: Hasenpusch (1977), Normandeau (1979) and Hartnagel (1978).

4. Secondary to the last remark, such a package should be widely available and generally usable in the social science research community. There is no question that SPSS satisfies these criteria, together with comments (2) and (3) above.

UCR and Microdata

The traditional view of the UCR data is presented by Oosthoek (1978). Citing this report:

> There are two dimensions to the Uniform Crime Report (UCR) CATEGORIES. The first consists of offence categories such as described in 2.2.2. The second dimension consists of the variables within each of those main categories. There are 10 such variables:
>
> 1. Total offences reported or known to police.
> 2. Total unfounded offences.
> 3. Total actual offences reported.
> 4. Total offences cleared by charge.
> 5. Total offences cleared "otherwise".
> 6. Total adult males charged.
> 7. Total adult females charged.
> 8. Total juvenile males charged.
> 9. Total juvenile females charged.
> 10. Total juveniles informally dealt with (new in 1974). (Oosthoek, 1978:32)

As may be seen from this list of "variables", each category refers to a total. What may be argued is that Oosthoek's "second dimension" refers to at least three independent dimensions, which may be taken as 3 levels of microdata, these being:

Type of Offence:	Unfounded Actual
Type of clearance:	Cleared by charge Cleared otherwise
Type of offender:	Adult males Adult females Juvenile males Juvenile females Juvenile informal

Additionally, it is possible to partially disaggregate offender type into two nominal variables -- age and sex.

Viewed from the perspective that one incident of crime is the basic unit of analysis, this present arrangement of the data, as described by Oosthoek, poses problems for most conventional statistical packages. The reason is outlined below.

The Data Matrix

SPSS assumes a rectangular data matrix such as is illustrated in Diagram 1. Each row represents all the information for one unit or case and each column represents the value of one variable for this unit or case.*

* It is important to remember that incidents are recorded with only one offence each being counted per incident (the most serious, according to MSO rules), even when there may be more than one offence. Thus the data do not and cannot permit an enumeration of offensive behaviour but only of offensive incidents which are incompletely characterized by the recording of the most serious offence.

DIAGRAM 1

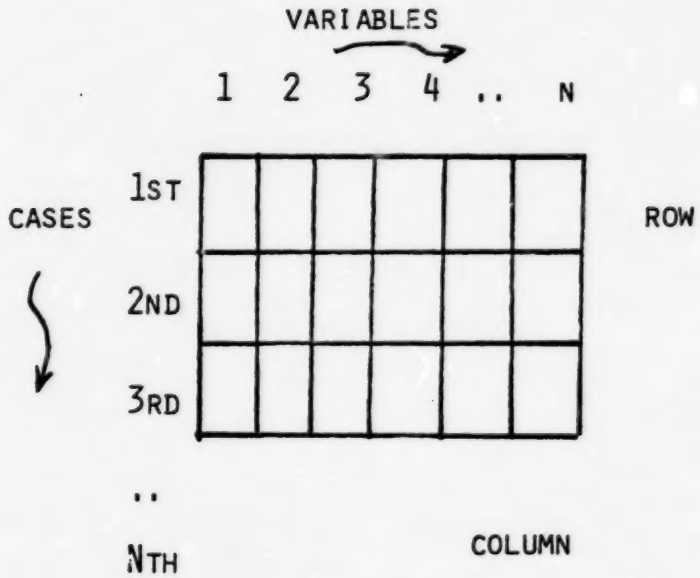


DIAGRAM 2

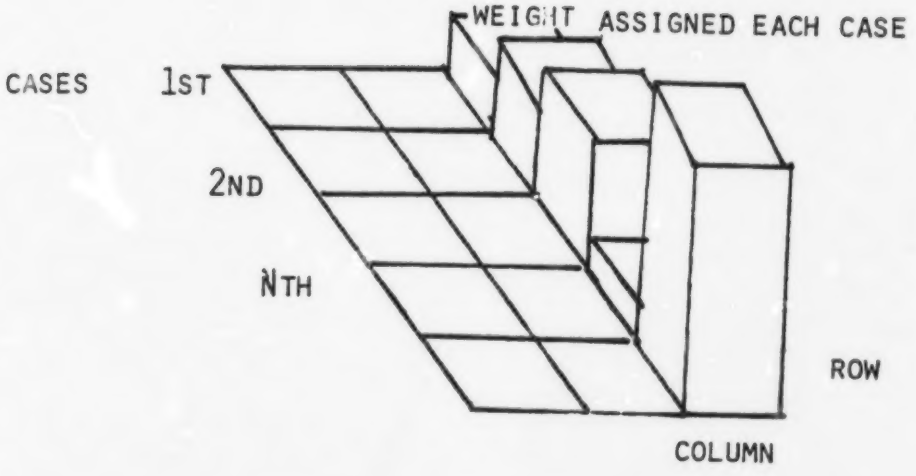
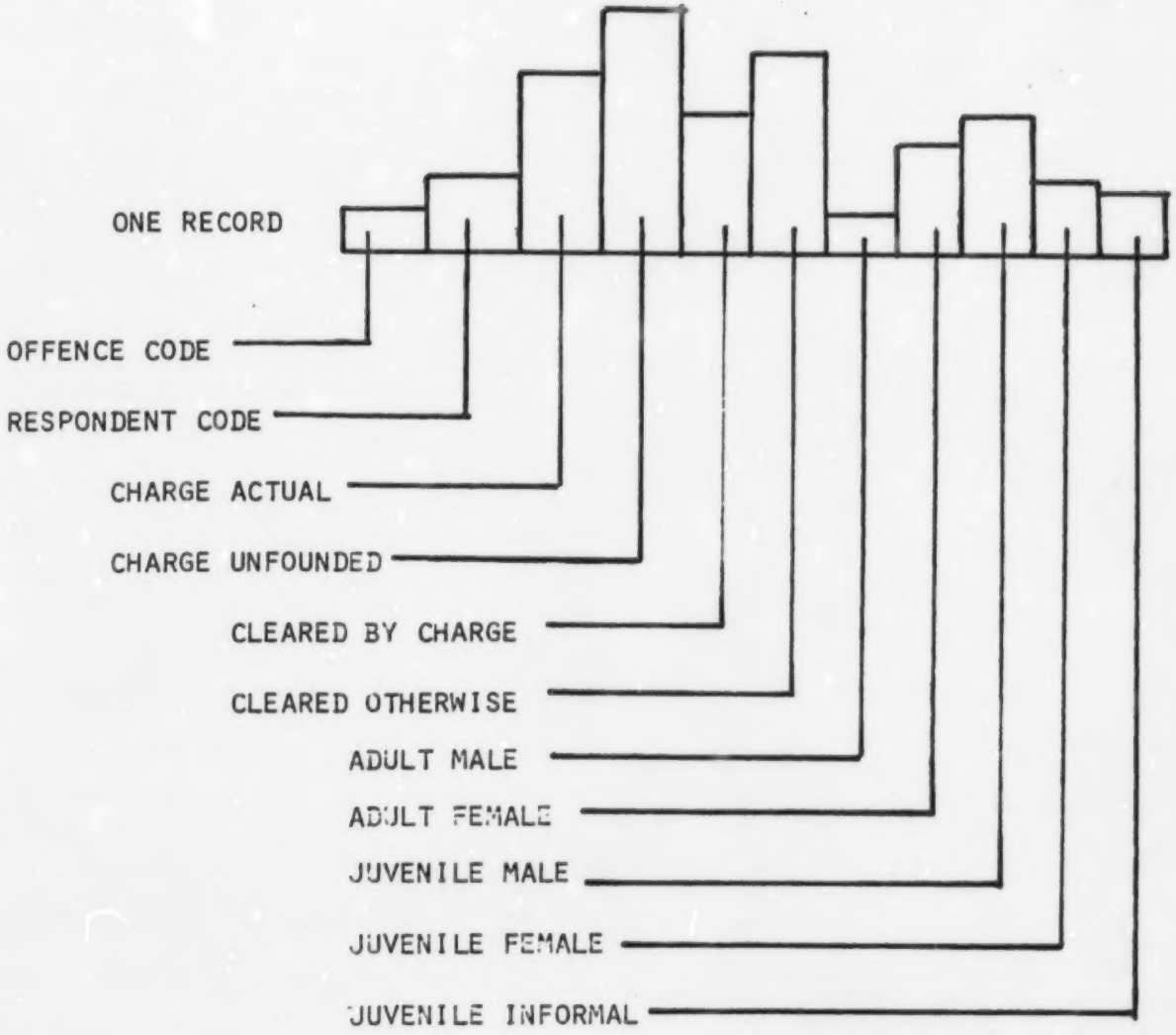


DIAGRAM 3


ONE RECORD
OFFENCE CODE
RESPONDENT CODE
CHARGE ACTUAL
CHARGE UNFOUNDED
CLEARED BY CHARGE
CLEARED OTHERWISE
ADULT MALE
ADULT FEMALE
JUVENILE MALE
JUVENILE FEMALE
JUVENILE INFORMAL

This assumption is inconvenient, however, when one or more column values of the matrix represents a total. In other words, this occurs when the data matrix presents an aggregated table and the row no longer represents one unit or case, but rather a cell frequency which is described by the other columns (variables) of the matrix in terms of the categorical values of the table of which it is a component. This situation is represented in Diagram 2.

Perhaps the most frequent instance of this situation in the social sciences occurs when the data are derived from a sample survey and each case is weighted to satisfy some sampling criterion.

Such a situation is directly tractable in SPSS, as it is in most statistical packages, by denoting the column containing the cell frequency as a WEIGHT factor.

The UCR data, however, have an additional wrinkle in that each record or case (a line in an aggregated reporting form) contains 3 independent univariate tables. One of these records is depicted in Diagram 3. While the presence of the three independent tables does not pose a problem*, current

* To deal with this we chose a subfile structure, chiefly for user convenience, as it entails fewer control cards. There may also be marginal savings in processing time.

limitations of SPSS, and to our knowledge most other packages, dictate that you cannot simultaneously weight more than one variable per case.

In this sense, traditional tabulation would miss a level of microdata, namely each particular dimension under consideration (for example, offender type). While some packages do permit the presentation of several variables across levels of other variables we would stress that this latter feature is intended to be for display purposes only, rather than for analytical purposes as it is here.

The strategy we outline is intended to convert Diagram 3 to Diagram 2, which does fully exploit the level of microdata inherent in the UCR data matrix. The result of the manipulations is the most generally applicable data configuration possible, a type of configuration which is neither microdata nor aggregate data per se (it is a hybrid of the two). It is readily transparent that the procedure also has extensions to the analysis of user summary tapes, which are increasing in popularity as a mode of data dissemination. One could just as well take published hard copy tables and analyze the implict microdata therein.*

* Assuming a reasonable number of dimensions and categories in tables. We might also point out that SPSS now has a REPORT feature which production of a custom designed product which could be made to order for each application or design convention.

The Procedure

While this data base may be viewed from several analytical perspectives, the sequence of events outlined below was chosen as being one method permitting considerable flexibility within the SPSS program.*

Assumptions:

1. The basic unit of measurement is a recorded instance of crime.
2. For demonstration purposes, it will suffice to have one annual record for each offence code.**
3. As they are recorded separately, the user will wish to consider the distributions on the following three variables separately:

    Type - unfounded

        - actual

    Clearance - cleared by charge

        - cleared otherwise

---

* Page references are made, where appropriate, to the second edition of Statistical Package for the Social Sciences.

* Only if the offence code is reported at least once during the calendar year.

Offender type - male adult
- female adult
- male juvenile
- female juvenile
- juvenile informal

4. The values of the variables: respondent type, population, and group size remain constant for each reporting unit during the year.

Step One - Preliminary

The creation of the SPSS system file that will satisfy the criteria listed above must be preceded by two intermediate SPSS jobs. These are:

1. A job using the AGGREGATE subprogram which condenses the monthly data into one annual record for each offence code for each reporting unit.

2. A job which reads the binary output file created by the AGGREGATE subprogram and subsequently uses the WRITE CASES subprogram to write the data into three subfiles with the annual offence totals written in the same column of the data matrix.*

* This is not really necessary providing that SPSS has information for each case indicating where the total is written.

Details of above job steps

A. AGGREGATE Subprogram (pp.206-17)

The purpose of this job is to compile a summary record for each offence code for each reporting unit. Within a year, there are potentially but not necessarily 12 monthly reports (records) to be collapsed into one summary record. This program permits considerable choice regarding the summarization of other variables in the data set on the output file (p. 208). For the sake of this demonstration we require an annual total for all offence codes to appear on the output file; and as we have assumed that respondent type, population size, and group size remain constant within a year, we will wish to have the mean value of these variables (which should be a constant) written on the output file.* (If it were desirable to have a record of the number of months of data included in the yearly total, then a dummy variable would be written on the summary record for each reporting unit).

* This assumes no bad data. One could check this assumption by also writing out either the minimum or maximum data in addition to the mean and comparing them on a separate job. It could also be printed on voluminous amounts of paper.

The procedure for doing this is as follows. The raw data is first read in the normal way using a standard input format. Then the data file must be sorted with the SORT CASES subprogram (pp.160-64) so that all similar offence codes within each respondent code are sequential. The AGGREGATE subprogram then follows:

```
COMPUTE      M = 1
SORT CASES   V1, V4 (A)
AGGREGATE    GROUPVARS = V1, V4 /
             VARIABLES  =  M,V5,V6,V7,V8,V9,V10,V11,V12,
                           V13,V14 /
             AGGSTATS = SUM /
             VARIABLES  =  V3,V15,V16,V17  /  AGGSTATS  =
                           MEAN /
             OPTIONS      3
             STATISTICS   ALL
```

From the printed record of this job it is important to retain the summary of the contents of the output file. This summary indicates the locations of the output variables on the binary file and also prints the number of records on the output file. This information is required for the next run.

The printed record looks like the following example:

```
AGGREGATION GROUP NUMBER
TOTAL NUMBER OF CASES IN AGGREGATION GROUP
V1      - GROUPING VARIABLE VALUE
V4      - GROUPING VARIABLE VALUE
M       - SUM
V5      - SUM
V6      - SUM
V7      - SUM
V8      - SUM
V9      - SUM
V10     - SUM
V11     - SUM
V12     - SUM
V13     - SUM
V14     - SUM
V3      - MEAN
V15     - MEAN
V16     - MEAN
V17     - MEAN
```

A TRUE AGGREGATED OUTPUT FILE WAS WRITTEN ON LOGICAL UNIT 9. IT CONTAINS 95156 AGGREGATED CASES. THE VARIABLES ON THE FILE GENERATED CAN BE READ INTO SPSS IN A SUBSEQUENT RUN USING THE "BINARY" FORMAT.

B. WRITE CASES Subprogram

This job is necessitated by the way in which SPSS processes variables that represent more than one occurrence of an event. SPSS does this by recognizing the value of a WEIGHT factor (pp. 129-31). If, for example, a variable is declared as a WEIGHT variable, and in a particular case that value is 6, the case will be treated in subsequent statistical calculations as if it had occurred 6 times. The

difficulty with the UCR data, however, is that multiple totals occur for each record and it would be difficult to devise a system for double or triple weighting each case.* The method chosen to circumvent this difficulty is to identify the categories of the three groups of totals: type, clearance, and offender code, as variables on three separate subfiles. These variables are written in three separate fields, one for each subfile, to permit the incorporation of labelling information in the final SPSS system file.

The first step of this job is to read back the binary data file that was created by the AGGREGATE subprogram (p. 47). Then dummy variables are created which serve to represent the category values of the three variables mentioned above. Finally, the data file is written out <u>three</u> times, once for each variable: type, clearance, and offender code (pp. 139-143).

```
INPUT MEDIUM        TAPE
# OF CASES          95156
VARIABLE LIST       AGG,TOTAL,V1,V4,M,V5 to V14,V15,V16,
                    V17
INPUT FORMAT        BINARY
COMPUTE             D=0
COMPUTE             F1=1
COMPUTE             F2=2
COMPUTE             F3=3
COMPUTE             F4=4
COMPUTE             F5=5
```

* Under present constraints of the program. For example, SPSS weighting is non-cumulative, thus ruling out a double-weighting strategy.

```
WRITE CASES        (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0 /
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,F1,D,D,V6,
                   V1,V4,V3,M,V15,V16,V17,F2,D,D,V7
WRITE CASES            (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0/
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,D,F1,D,V8,
                   V1,V4,V3,M,V15,V16,V17,D,F2,D,V9
WRITE CASES        (4(F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0 /)
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,D,D,F1,V10,
                   V1,V4,V3,M,V15,V16,V17,D,D,F2,V11,
                   V1,V4,V3,M,V15,V16,V17,D,D,F3,V12,
                   V1,V4,V3,M,V15,V16,V17,D,D,F4,V13,
                   V1,V4,V3,M,V15,V16,V17,D,D,F5,V14,
```

At this point, a data file has been created which will permit tabulation considering an instance of crime as the case. This file will have a subfile structure that will permit the separate treatment of type, clearance, and offender code.

Step Two - Final System File

The first step in the creation of this file is the declaration of the subfile structure (p.39). For each subfile the number of cases is determined by multiplying the number of summary records (from AGGREGATE subprogram summary print) by the number of categories in the chosen sub-file designating variables (eg. Type=2; Clear=2; and offender code=5). Thus in the TYPE subfile, the complete information was written out once for each of the unfounded and actual number categories. Therefore the number of cases in the subfile TYPE is 95156 x 2 = 190312. The VARIABLE LIST and

INPUT FORMAT control cards are written according to the layout on the WRITE CASES card of the previous run. Note that the offence total is declared as the WEIGHT variable. The variables FLAG1, FLAG2, and FLAG3 are the tabulating variables for each subfile. Thus if one wished to tabulate clearance by offence code one would run the relevant subfile (CLEAR) and request a table; offence code by FLAG3.

Creating the final file

```
SUBFILE LIST      TYPE (190312)
                  CLEAR (190312)
                  CULPRIT (475780)
INPUT MEDIUM      TAPE
VARIABLE LIST     RCODE,OFFCODE,YEAR,MONTH,PTYPE,PSIZE,
                  GSIZE,FLAG1,FLAG2,FLAG3,WEIGHT
INPUT FORMAT      FIXED (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
WEIGHT            WEIGHT
SELECT IF         (WEIGHT GT 0)
VAR LABELS        FLAG1,TAB CHARACTERISTIC FOR SUBFILE1/
                  FLAG2,TAB CHARACTERISTIC FOR SUBFILE2/
                  FLAG3,TAB CHARACTERISTIC FOR SUBFILE3/
VALUE LABELS      FLAG1,(0) INAPPLICABLE (1) CLEAR CHARGE
                  (2)ACTUAL/FLAG2(0)INAPPLICABLE(1)
                  (1)CLEAR CHARGE
                  (2)CLEAR OTHER/FLAG3(0)INAPPLICABLE(1)
                  MALE ADULT
                  (2)FEMALE ADULT(3)MALE JUVENILE(4)
                  FEMALE JUVENILE(5)JUVENILE INFORMAL/
LIST FILEINFO     COMPLETE
SAVE FILE         SFINAL
```

Additional processing

There is an additional variable that could be created by the user. One might wish to break down offender code by sex and age. This cannot be completely done as the juvenile informal category is not distributed by sex. What could be done is the following:

```
RUN SUBFILES    (CULPRIT)
COMPUTE         SEX = FLAG3
RECODE          SEX (1,3=1)(2,4=2)(5=0)
COMPUTE         AGE  = FLAG3
RECODE          AGE  = (1,2=1)(3,4,5=2)
VALUE LABELS    SEX (1)MALE(2)FEMALE(0)MISSING/
                AGE (1)ADULT(2)JUVENILE
```

REFEFENCES

Timothy F. Hartnagel, "The Effect of Age and Sex Compositions of Provincial Populations on Provincial Crime Rates". Canadian Journal of Criminology, 20, 1, 1978, 28-33.

Burkhard Hasenpusch, "The Rise and Fall of Crime in Canada - An Attempt at Criminological Froecasting". Unpublished manuscript, 1977.

James A Inciardi, "The Uniform Crime Reports: Some Considerations of their shortcomings and utility". Review of Data Use, 6, (November, 1978) 3-16.

Andre Normandeau, Daniel Elie, and Burkhard Hasenpusch, "Forecasting the Future Development of Policing in Canada". Université de Montréal, École de Criminologie: Final Report on the Research Grant No. 410-770186, Jan. 1979.

A. Oosthoek, The Utilization of Official Crime Data. Ottawa: Minister of Supply and Services Canada, 1978.

Statistics Canada / Statistique Canada

Research Development and Analysis Group

Section de la recherche du développement et de l'analyse

R-018
C.1

TRANSFORMATION DES DONNÉES DU PROGRAMME DE DÉCLARATION UNIFORME DE LA CRIMINALITÉ EN FONCTION DES DONNÉES SUR LES CAS

Document de recherche No. 1

par

Craig McKie

La responsabilité
de l'analyse et de l'interprétation des données
incombe à l'auteur et non à Statistique Canada

Direction de la statistique des institutions et de l'agriculture Section de la recherche, du développement et de l'analyse Statistique Canada

1981

La tenue des dossiers administratifs pose un problème général qui ne devient manifeste qu'au moment de l'interprétation ultérieure des fichiers de données sommaires. Ce problème découle du fait que les organismes, lorsqu'ils accumulent des dossiers sur divers phénomènes, peuvent créer par inadvertance pour eux-mêmes des structures de données qui sont difficiles à manipuler. Les programmeurs se sont familiarisés avec les matrices de données rectangulaires et les façons de les manipuler, mais les difficultés que pose la manipulation des matrices non rectangulaires (qui se trouvent souvent dans les dossiers administratifs où l'on tient des totaux sommaires) en ont restreint dans une grande mesure l'utilisation. Dans ces cas, il arrive souvent que les données soient considérées comme étant difficiles à manipuler, sauf dans le cas des descriptions élémentaires. Rarement des systèmes coûteux sont-ils établis sur demande pour transformer ces données. Ces systèmes sont coûteux, peuvent nécessiter la connaissance d'un langage de programmation, peuvent être assujettis à des contraintes d'installation et peuvent ne pas être assez documentés pour survivre au départ de leur créateur ou se prêter facilement à des modifications.

Le présent document présente une technique fondée sur le progiciel statistique des sciences sociales qui permet de résoudre ces problèmes grâce à une méthode généralisée de manipulation de matrices de données non rectangulaires comme celles qu'on trouve dans les dossiers administratifs sommaires. Nous avons établi cette technique à la Division de la statistique judiciaire de Statistique Canada pour accroître la gamme des manipulations statistiques possibles des données globales sur la criminalité et l'application de la loi produites dans le cadre du programme de déclaration uniforme de la criminalité. C'est pourquoi

nous utilisons, à titre d'exemple, les données du programme de déclaration uniforme de la criminalité.

Les ensembles de données du genre de celles qui proviennent du programme de déclaration uniforme de la criminalité, où les totaux mensuels sur un grand nombre de genres d'incidents criminels sont saisis de façon électronique, sont limités et susceptibles de poser des problèmes si l'on veut procéder à des analyses de tableaux qui ne soient pas rudimentaires. Ces problèmes découlent de la nature périodique et sommaire des données recueillies.

Données du programme de déclaration uniforme de la criminalité

Les données du programme DUC sont recueillies par un certain nombre d'organismes (notamment par Statistique Canada et par le FBI aux États-Unis) à partir d'un grand nombre d'unités de déclaration, qui correspondent généralement aux territoires des corps de police municipaux, régionaux et provinciaux (il y en a environ 1,800 au Canada). Ces dossiers sont "la seule source de données sur l'ampleur et les tendances de la criminalité.... et les administrateurs, les hommes politiques, les preneurs de décisions, la presse et le grand public y ont largement recours" (Inciardi, 1978:3). Cependant en raison de leur structure et de l'absence de contrôle de la qualité au moment de leur établissement*, les rapports d'information fondés sur ces données sont généralement considérés comme "incomplets et entachés de biais, d'où la naissance et la persistance de nombreux

* On croit généralement que de nombreux crimes ne sont jamais déclarés à la police; il y a peut-être 5 cas non déclarés pour chaque cas déclaré.

mythes au sujet de la criminalité (1978:3). Nous ne pouvons rien faire ici au sujet de ce dernier genre de problème** (déclarations faites au hasard, biaisée ou erronées), mais il est possible d'apporter une solution technique au premier problème concernant les limites du dépouillement statistique.

Jusqu'à présent, la nature globale des données a restreint dans une grande mesure l'analyse de ces données. Des totaux sont recueillis et additionnés, mais on ne tente pas de tenir un registre de chacun des incidents ou des contrevenants (lorsqu'on connaît leur identité). On additionne plutôt, souvent une fois par mois, les incidents de divers genres déclarés à la police qui les signale par la suite pour permettre d'établir les totaux à l'échelle nationale. Il n'en résulte que des totalisations globales -- et il existe peu d'exemples dans la documentation canadienne ou américaine où l'on ait conçu des moyens plus ingénieux pour analyser ce genre de données.

La solution consiste donc à remanier les données du programme DUC de façon à pouvoir procéder à une analyse plus rigoureuse. Pour ce faire, il s'agit uniquement de savoir comment mieux traiter et manipuler les données actuelles.

** Les données du programme DUC comportent une foule de difficultés. Elles sont insuffisantes en ce qui concerne: l'exhaustivité des déclarations, la précision, la fiabilité et l'uniformité de la classification des infractions; et l'application de la règle de l'"infraction la plus grave" qui se traduit par l'absence dans les données des accusations moins graves mais connexes portées en même temps que les infractions déclarées.

La problématique

La problématique des données administratives du programme DUC du Canada est la suivante: il s'agit d'un vaste ensemble de tableaux mensuels où figurent des chiffres sommaires sur un grande nombre de catégories d'infractions (132). Chaque dossier représente donc l'activité totale mensuelle d'une unité de déclaration pour chacun des 132 genres d'infraction.

Nous essaierons ci-dessous de montrer comment ces données peuvent être remaniées et transformées pour permettre de produire un fichier "SAVE FILE" du SPSS qui peut produire des tableaux annuels type.

Fusion du programme DUC et du système SPSS

La meilleure façon d'expliquer le fonctionnement de ce programme (et l'on épuise rapidement le vocabulaire traditionnel pour y arriver) est peut-être de dire qu'il fonctionne comme si l'on se servait de microdonnées détaillées dans le cadre du SPSS. On pourrait aussi l'expliquer de deux autres façons: a) les données peuvent être utilisées de concert avec le SPSS d'une manière assez traditionnelle une fois que le traitement intermédiaire est terminé ou b) la structure des données est remaniée de telle sorte qu'elle se comporte comme des microdonnées (le cas échéant).

Comme c'est le cas, il est, à strictement parler, trompeur de faire état de la création de microdonnées lorsqu'il n'existe que des données globales. Il faudrait plutôt se demander, nous semble-t-il, pourquoi il faudrait utiliser le

SPSS de cette façon pour dépouiller ces chiffres. Il y a un certain nombre de raisons:

1. Le SPSS est en voie d'être le progiciel le plus universellement appliqué. Le produit permettrait donc beaucoup plus à des profanes de le manipuler, ce que ne permet pas le "système" actuel qui fait appel à un certain nombre de programmes distincts faits sur mesure dont la portée est extrêmement limitée.

2. L'utilisation du SPSS ne crée pas au sens strict de nouvelles microdonnées. Elle permet plutôt de réaménager les données en vue d'exploiter au maximum le niveau de microdonnées qui est implicite dans le présent dossier sommaire et dans d'autres dossiers globaux qui sont stockés sous forme de matrice. (Voir texte ci-dessous).

3. Comme c'est le cas pour tous les domaines spécialisés de la statistique sociale, l'étude de la criminologie devient rapidement plus complexe; c'est pourquoi une grande variété de méthodes statistiques à variables multiples sont appliquées aux données sur la criminalité. Certaines de ces applications sont fondées sur les données du programme DUC comme: les catégories d'infraction (Oosthoek, 1978), et l'analyse des tendances (Inciardi, 1978,) qui s'applique logiquement à l'analyse des séries

---

* Pour d'autres exemples du Canada, voir également: Hasenpusch (1977), Normandeau (1979) et Hartnagel (1978).

chronologiques*. A la lumière de ces approches récentes, la nécessité de pouvoir manipuler les données du programme DUC devient évidente, surtout lorsqu'il faut utiliser des progiciels statistiques (du point de vue pratique) pour avoir recours à ces méthodes.

4. Pour faire suite à la dernière observation, un progiciel de ce genre doit être disponible à tous et être susceptible d'être utilisé par les chercheurs des sciences sociales. Il ne faut aucun doute que le SPSS répond à ces critères tout comme les observations en (2) et (3) ci-dessus.

Programme DUC et microdonnées

Oosthoek présente le point de vue traditionnel sur les données du programme DUC. Voici des extraits de son étude:

> Les CATÉGORIES du programme de déclaration uniforme de la criminalité présentent deux dimensions. La première consiste dans les catégories d'infraction comme celles dont il a été fait état en 2.2.2. La deuxième englobe les variables de chacune de ces catégories principales. Il y a 10 variables de ce genre:
>
> 1. Les infractions communiquées à la police ou connues d'elle.
> 2. Les infractions non fondées.
> 3. Les infractions réelles déclarées.
> 4. Les infractions classées par mise en accusation.
> 5. Les infractions classées sans mise en accusation.
> 6 Le nombre d'hommes inculpés.
> 7. Le nombre de femmes inculpées.

8. Le nombre de jeunes de sexe masculin inculpés.
9. Le nombre de jeunes de sexe féminin inculpés.
10. Le nombre de jeunes traités hors cours (nouvelle catégorie introduite en 1974. (Oosthoek, 1978:32).

Comme en témoigne cette liste de "variables", chaque catégorie se rapporte à un nombre total. On peut soutenir que la "deuxième dimension" d'Oosthoek a trait à au moins trois dimensions indépendantes, qui peuvent être considérées comme 3 niveaux de microdonnées, soit:

Genre d'infraction:	Non fondée
	Réelle
Genre de classement:	Classée par mise en accusation
	Classée sans mise en accusation
Genre de contrevenant:	Hommes
	Femmes
	Jeunes de sexe masculin
	Jeunes de sexe féminin
	Jeunes traités hors cours

En outre, il est possible de séparer partiellement le genre de contrevenant en deux variables: l'âge et le sexe.

Comme l'unité de base d'analyse est un incident criminel, la structure des données décrite par Oosthoek pose des difficultés pour la plupart des progiciels statistiques traditionnels. La raison en est donnée ci-dessous.

La matrice de données

Le SPSS suppose un matrice de données rectangulaires comme celle qui figure au diagramme 1. Chaque rangée représente tous les renseignements d'une unité ou d'un cas et chaque colonne représente la valeur d'une variable de cette unité ou de ce cas.*

* Il importe de ne pas oublier qu'on ne compte qu'une infraction par incident déclaré (la plus grave, selon la règle de l'infraction la plus grave), même lorsqu'il peut y avoir plus d'une infraction. Ainsi les données ne permettent pas de décrire le comportement criminel, mais seulement les incidents criminels qui sont décrits d'une façon incomplète par l'enregistrement de l'infraction la plus grave.

DIAGRAMME 1

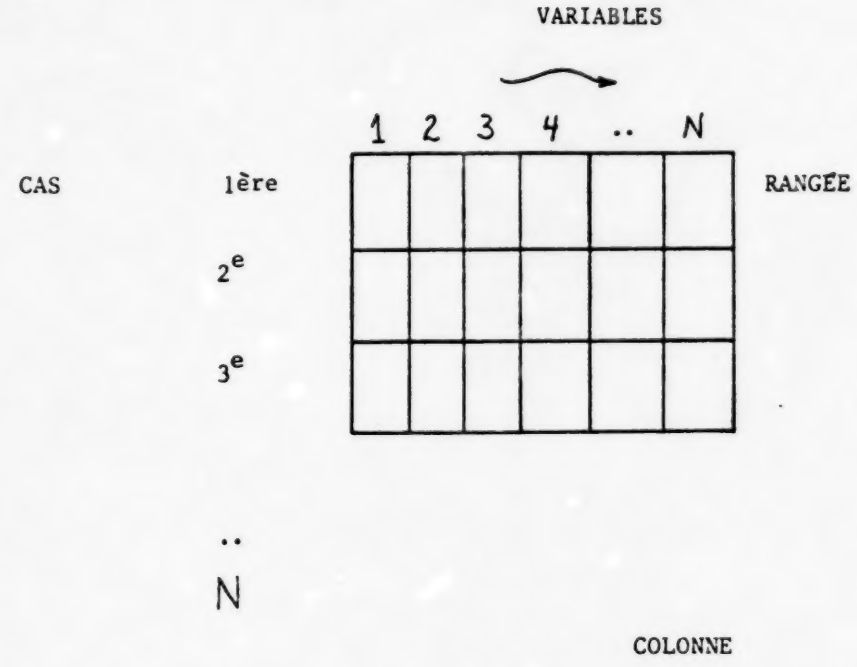


DIAGRAMME 2

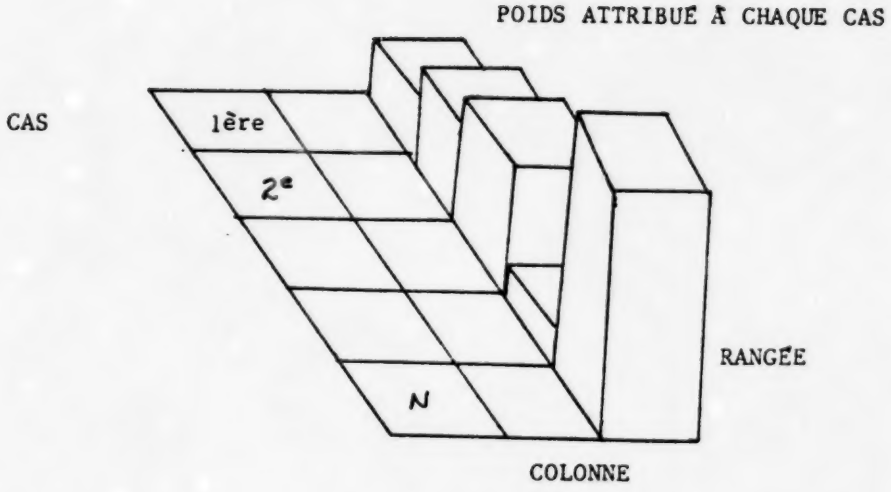
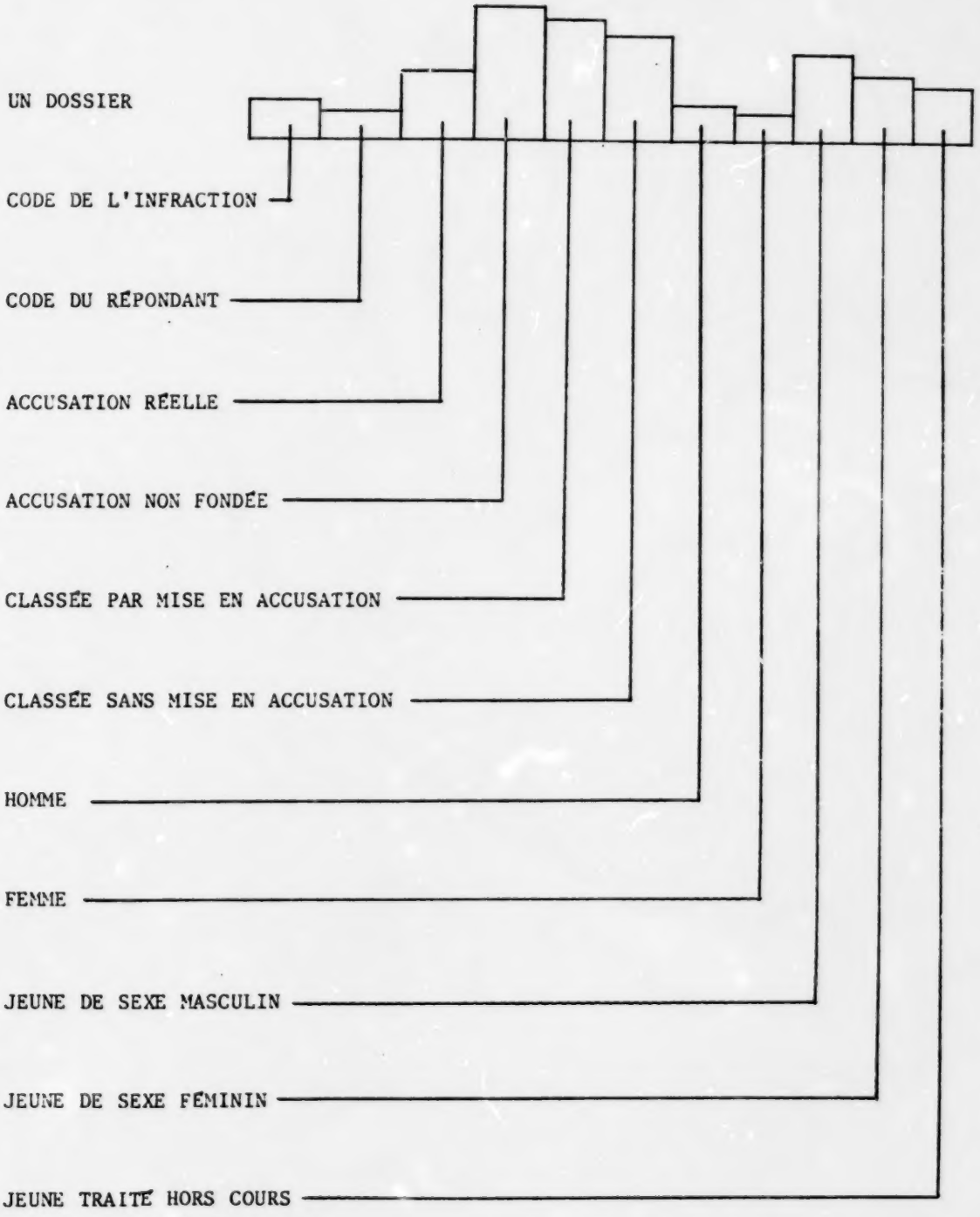


DIAGRAMME 3


UN DOSSIER
CODE DE L'INFRACTION
CODE DU RÉPONDANT
ACCUSATION RÉELLE
ACCUSATION NON FONDÉE
CLASSÉE PAR MISE EN ACCUSATION
CLASSÉE SANS MISE EN ACCUSATION
HOMME
FEMME
JEUNE DE SEXE MASCULIN
JEUNE DE SEXE FÉMININ
JEUNE TRAITÉ HORS COURS

Cette hypothèse est cependant peu pratique lorsqu'une ou plusieurs valeurs des colonnes de la matrice représentent un total. En d'autres mots, cela se produit lorsque la matrice de données présente un tableau global et que la rangée ne représente plus une unité ou un cas, mais un effectif de classe qui est décrit par les autres colonnes (variables) de la matrice en fonction des valeurs des catégories du tableau dont elle fait partie. Cette situation est représentée au diagramme 2.

Cette situation se présente peut-être le plus souvent dans les sciences sociales lorsque les données proviennent d'une enquête par échantillon et qu'on pondère chaque cas pour répondre à un critère d'échantillonnage.

Il est possible de faire face à une situation de ce genre dans le SPSS, comme dans la plupart des progiciels statistiques, en attribuant à la colonne contenant un effectif de clase un facteur de PONDÉRATION.

Les données du programme DUC ont cependant un élément supplémentaire, car chaque dossier ou cas (une ligne dans une formule de déclaration globale) contient 3 tableaux indépendants à une seule variable. L'un de ces dossiers figure au diagramme 3. La présence des trois tableaux indépendants ne pose pas de problème*, mais les limites actuelles du SPSS et, à notre connaissance, de la plupart des autres progiciels, interdisent de pondérer simultanément plus d'une

* Pour faire face à cette situation, nous avons choisi une structure de sous-fichier, principalement pour les besoins des utilisateurs, car il comporte moins de cartes de contrôle. On peut également épargner du temps de traitement.

variable par cas.

En ce sens, les totalisations traditionnelles ne tiendraient pas compte d'un niveau de microdonnées, à savoir chaque dimension particulière à l'étude (par exemple, le genre de contrevenant). Bien que certains progiciels permettent de présenter plusieurs variables par rapport à d'autres variables, nous insistons sur le fait que cette dernière caractéristique ne sert qu'aux fins de démonstration et non pas aux fins d'analyse comme c'est le cas ici.

La méthode que nous exposons vise à transposer le diagramme 3 dans le diagramme 2, qui exploite totalement les microdonnées de la matrice de données du programme DUC. Les manipulations permettent de présenter les données de la façon la plus applicable possible: les données ne sont ni des microdonnées ni des données globales (il s'agit d'un hybride des deux). Il est évident que la méthode peut s'appliquer à l'analyse des bandes sommaires des utilisateurs, qui gagnent en popularité comme mode de diffusion des données. On pourrait en outre prendre des tableaux publiés en clair et analyser les microdonnées implicites qui s'y trouvent.*

---

* A supposer qu'il y ait un nombre raisonnable de dimensions et de catégories dans les tableaux. Nous pourrions également signaler que le SPSS a une caractéristique REPORT qui permet de produire sur demande des données qui pourraient servir pour chaque application ou présentation.

La façon de procéder

La présente base de données peut être analysée selon plusieurs points de vue, mais l'ordre des événements décrits ci-dessous a été choisi parce qu'il permettait beaucoup de souplesse dans le cadre du programme SPSS.*

Hypothèses:

1. L'unité de mesure de base est un incident criminel déclaré.
2. Aux fins de démonstration, il suffira d'avoir un dossier annuel pour chaque code d'infraction.**
3. Comme les trois variables suivantes sont déclarées séparément, l'utilisateur voudra examiner séparément la distribution de chacune d'elles:

   Genre - non fondée
   - réelle

   Classement - classée par mise en accusation
   - classée sans mise en accusation

---

* Les pages renvoient, s'il y a lieu, à la deuxième édition du Progiciel statistique des sciences sociales.

** Seulement si le code de l'infraction est déclaré au moins une fois au cours de l'année civile.

Genre de contrevenant - homme

femme

jeune de sexe masculin

jeune de sexe féminin

jeune traité hors cours

4. Les valeurs des variables: le genre de répondant, la population et la taille du groupe restent constants pour chaque unité de déclaration au cours de l'année.

Première étape - Préliminaire

La création du fichier du système SPSS qui satisfera aux critères susmentionnés doit être précédée de deux étapes intermédiaires, notamment:

1. Une étape utilisant le sous-programme AGGREGATE qui condense les données mensuelles en un dossier annuel pour chaque code d'infraction de chaque unité de déclaration.

2. Une étape qui lit le fichier sortie binaire créé par le sous-programme AGGREGATE et utilise par la suite le sous-programme WRITE CASES pour écrire les données dans trois sous-fichiers où les totaux des infractions annuelles figurent dans la même colonne de la matrice de données.*

* Il n'est pas réellement nécessaire de procéder ainsi si le SPSS contient des renseignements pour chaque cas indiquant où le total est écrit.

Détail des étapes ci-dessus

A. Sous-programme AGGREGATE (pp. 206-217)

Il s'agit d'établir un dossier sommaire pour chaque code d'infraction de chaque unité de déclaration. Au cours d'une année, il peut y avoir, mais pas nécessairement, 12 rapports mensuels (dossiers) à assembler dans un dossier sommaire. Ce programme permet un vaste choix en ce qui concerne l'établissement d'un résumé d'autres variables de l'ensemble de données du fichier sortie (p. 208). Pour les besoins de cette démonstration, nous avons besoin du total annuel de tous les codes d'infraction qui doivent figurer dans le fichier sortie; et comme nous avons supposé que le genre de répondant, la taille de la population et la taille du groupe restent constants au cours d'une année, nous désirerons que la valeur moyenne de ces variables (qui doit être constante) figure dans le fichier sortie.* (S'il était souhaitable que le nombre de mois pour lesquels on dispose de données figure dans le total de l'année, une variable vide serait inscrite dans le dossier sommaire de chaque unité de déclaration).

---

* A supposer qu'il n'y ait pas de mauvaises données. On pourrait vérifier cette hypothèse en écrivant les données minimales ou maximales en plus de la moyenne et en les comparant dans un travail distinct. On pourrait aussi l'imprimer sur de grandes quantités de papier.

Voici la marche à suivre. Les données brutes sont d'abord lues de la façon normale à l'aide d'une présentation type. Le fichier de données doit ensuite être trié avec le sous-programme SORT CASES (p. 160-164) pour que tous les codes d'infraction semblables de chaque code de répondant soient consécutifs. Le sous-programme AGGREGATE est le suivant:

```
COMPUTE          M = 1

SORT CASES       V1, V4 (A)

AGGREGATE        GROUPVARS = V1, V4 /
                 VARIABLES = M, V5, V6, V7, V8, V9, V10, V11, V12, V13,
                             V14/

                 AGGSTATS = SUM /
                 VARIABLES = V3, V15, V16, V17 / AGGSTATS =
                             MEAN /
                 OPTIONS     3

                 STATISTICS   ALL
```

D'après le dossier imprimé de ce travail, il importe de retenir le sommaire du contenu du fichier sortie. Ce sommaire indique les emplacements des variables de sortie dans le fichier binaire et imprime également le nombre de dossiers dans le fichier sortie. Ces renseignements sont nécessaires pour la prochaine exécution.

Le dossier imprimé ressemble à ce qui suit:

```
AGGREGATION GROUP NUMBER
TOTAL NUMBER OF CASES IN AGGREGATION GROUP
V1  - GROUPING VARIABLE VALUE
V4  - GROUPING VARIABLE VALUE
M   - SUM
V5  - SUM
V6  - SUM
V7  - SUM
V8  - SUM
V9  - SUM
V10 - SUM
V11 - SUM
V12 - SUM
V13 - SUM
V14 - SUM
V3  - MEAN
V15 - MEAN
V16 - MEAN
V17 - MEAN
```

A TRUE AGGREGATED OUTPUT FILE WAS WRITTEN ON LOGICAL UNIT 9. IT CONTAINS 95156 AGGREGATED CASES. THE VARIABLES ON THE FILE GENERATED CAN BE READ INTO SPSS IN A SUBSEQUENT RUN USING THE "BINARY" FORMAT.

B. Sous-programme WRITE CASES

Il faut accomplir cette étape à cause de la façon dont le PSSS traite les variables qui représentent un événement qui se produit plus d'une fois. Le SPSS le fait en reconnaissant une variable comme facteur de pondération (WEIGHT) (pp. 129-131). Si, par exemple, une variable est déclarée comme variable de pondération (WEIGHT), et si dans un cas particulier cette valeur est 6, le cas sera traité dans les calculs statistiques ultérieurs comme si cela s'était produit 6 fois. La difficulté que présentent cependant les données DUC sont

qu'il y a des totaux multiples pour chaque dossier, et il serait difficile de concevoir un système permettant de pondérer deux ou trois fois chaque cas.* La méthode choisie pour surmonter cette difficulté consiste à identifier les catégories des trois groupes de totaux: genre, classement et code du contrevenant comme variables des trois sous-fichiers distincts. Ces variables sont écrites dans trois zones distinctes, une pour chaque sous-fichier, pour permettre d'incorporer les écritures de labels dans le fichier du système SPSS final.

La première étape de ce travail consiste à lire le fichier de données binaire créé par le sous-programme AGGREGATE (p. 47). Puis on crée des variables vides qui servent à représenter les valeurs des catégories des trois variables susmentionnées. Enfin, le fichier de données est écrit trois fois, une fois pour chaque variable: genre, classement et code du contrevenant (pp. 139-143).

```
INPUT MEDIUM        TAPE
# OF CASES          95156
VARIABLE LIST       AGG, TOTAL, V1, V4, M, V5 TO V14, V15, V16, V17
INPUT FORMAT        BINARY
COMPUTE             D = 0
COMPUTE             F1 = 1
COMPUTE             F2 = 2
COMPUTE             F3 = 3
COMPUTE             F4 = 4
COMPUTE             F5 = 5
```

* Selon les contraintes actuelles du programme. Par exemple, la pondération du SPSS n'est pas cumulative, ce qui exclut une stratégie de double pondération.

```
WRITE CASES        (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0 /
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,F1,D,D,V6,
                   V1,V4,V3,M,V15,V16,V17,F2,D,D,V7
WRITE CASES        (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0/
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,D,F1,D,V8,
                   V1,V4,V3,M,V15,V16,V17,D,F2,D,V9
WRITE CASES        (4(F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0 /)
                   F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
                   V1,V4,V3,M,V15,V16,V17,D,D,F1,V10,
                   V1,V4,V3,M,V15,V16,V17,D,D,F2,V11,
                   V1,V4,V3,M,V15,V16,V17,D,D,F3,V12,
                   V1,V4,V3,M,V15,V16,V17,D,D,F4,V13,
                   V1,V4,V3,M,V15,V16,V17,D,D,F5,V14,
```

A ce moment-ci, un fichier de données a été créé qui permettra la totalisation d'un incident criminel comme cas. Ce fichier aura une structure de sous-fichier qui permettra de traiter séparément le genre, le classement et le code du contrevenant.

Deuxième étape - Fichier du système final

La première étape de la création de ce fichier est la déclaration de la structure du sous-fichier (p. 39). Pour chaque sous-fichier, on détermine le nombre de cas en multipliant le nombre de dossiers sommaires (d'aprés l'imprimé sommaire du sous-programme AGGREGATE) par le nombre de catégories dans le sous-fichier choisi qui désigne les variables (par ex. genre = 5). Ainsi dans le sous-fichier du TYPE, les renseignements complets ont été écrits une fois pour chacune des catégories d'infraction non fondée et réelle. Par conséquent, le nombre du cas du sous-fichier TYPE est 95156 x 2 = 190312. Les cartes de contrôle VARIABLE LIST

et INPUT FORMAT sont écrites selon le schéma d'enregistrement de la carte WRITE CASES de l'exécution précédente. Il convient de noter que le total des infractions est déclaré comme variable de pondération (WEIGHT). Les variables FLAG1, FLAG2 ET FLAG3 sont les variables de totalisation pour chaque sous-fichier. Ainsi si l'on veut totaliser le classement selon le code de l'infraction, on exécute le sous-fichier pertinent (CLEAR) et on demande un tableau; code d'infraction par FLAG3.

Création du fichier final

```
SUBFILE LIST      TYPE (190312)
                  CLEAR (190312)
                  CULPRIT (475780)
INPUT MEDIUM      TAPE
VARIABLE LIST     RCODE, OFFCODE,YEAR,MONTH,PTYPE,PSIZE,
                  GSIZE,FLAG1,FLAG2,FLAG3,WEIGHT
INPUT FORMAT      FIXED (F5.0,F3.0,3F2.0,F5.0,4F1.0,F8.0)
WEIGHT            WEIGHT
SELECT IF         (WEIGHT GT 0)
VAR LABELS        FLAG1,TAB CHARACTERISTIC FOR SUBFILE1/
                  FLAG2,TAB CHARACTERISTIC FOR SUBFILE2/
                  FLAG3,TAB CHARACTERISTIC FOR SUBFILE3/
VALUE LABELS      FLAG1,(0) INAPPLICABLE (1) CLEAR CHARGE
                  (2) ACTUAL/FLAG2(0)INAPPLICABLE(1)
                  (1) CLEAR CHARGE
                  (2) CLEAR OTHER/FLAG3(0)INAPPLICABLE(1)
                  MALE ADULT
                  (2)FEMALE ADULT(3)MALE JUVENILE(4)
                  FEMALE JUVENILE(5)JUVENILE INFORMAL/
LIST FILEINFO     COMPLETE
SAVE FILE         SFINAL
```

Traitement additionnel

Il y a une autre variable que l'utilisateur peut créer. On pourrait désirer ventiler le code du contrevenant selon le sexe et l'âge. Il est impossible de le faire complètement, car la catégorie des jeunes traités hors cours n'est pas distribuée selon le sexe. Voici ce qu'on pourrait faire:

```
RUN SUBFILES        (CULPRIT)
COMPUTE             SEX = FLAG3
RECODE              SEX (1,3=1) (2,4=2)(5=0)
COMPUTE             AGE = FLAG3
RECODE              AGE = (1,2=1) (3,4,5=2)
VALUE LABELS        SEX (1) MALE (2) FEMALE (0) MISSING/
                    AGE (1) ADULT (2) JUVENILE
```

BIBLIOGRAPHIE

Timothy F. Hartnagel, "The Effect of Age and Sex Compositions of Provincial Populations on Provincial Crime Rates". Canadian Journal of Criminology, 20, 1, 1978, 28-33.

Burkhard Hasenpusch, "The Rise and Fall of Crime in Canada - An Attempt at Criminological Forecasting". Manuscrit non publié, 1977.

James A. Inciardi, "The Uniform Crime Reports: some Considerations of their shortcomings and utility". Review of Data Use, 6, (novembre 1978) 3-16.

André Normandeau, Daniel Elie et Burkhard Hasenpusch, "Forecasting the Future Development of Policing in Canada". Université de Montréal, Ecole de criminologie: rapport final sur la subvention de recherche nº 410-770186, janvier 1979.

A. Oosthoek, The Utilization of Official Crime Data. Ottawa: ministère des Approvisionnements et Services Canada, 1978.